RESEARCH PROPOSAL

Statistical Study of Predictive Analytic Techniques in Data Mining

A synopsis submitted to

SWAMI RAMANAND TEERTH MARATHWADA UNIVERSITY, NANDED (M.S.)

For the degree of

Doctor of Philosophy

In

STATISTICS

Under the Faculty of Science

By

Mr. Tangawade Atish Sitaram
M.Sc (Statistics)

Guided By

Dr. Aniket Avinash Muley
Assistant Professor & I/C Head,
Department of Statistics,
School of Mathematical Sciences,
Swami Ramanand Teerth Marathwada University,
Nanded-431606 (M. S.) INDIA

Table of Contents

Sr. No.	Contents	Page No.
1	Introduction	1
2	Review of Literature and Development in the subject	2
3	Objectives of the Research	3
4	Methodology	4
5	Importance of Study	5
6	Proposed Work Plan	7
7	References	8-9

Abstract:

Now-a-days, the amount of data stored in any database is increasing rapidly. As the world grows in complexity, overwhelming us with the data it generates. So development of computer field provide a framework for storage and such stored data is used in different areas such as business, medical, education, weather forecasting and climate change, share market, planning etc. this stored data is nothing but historical data for organization and which is very useful to answer to many questions related to day-to-day activities,strategic planning, the manual process of data analysis becomes tedious as size of data grows and number of dimensions increases, so the process of data analysis needs to be computerized. We also need discovering efficient solution and potentially useful information from a large amount of data and use it in efficient manner; the data mining becomes the only hope for elucidating the patterns that underlie it. Also, data mining is non-trivial process that automatically collects the useful hidden information from the data and is taken on as forms of rule, concept, pattern & so on.

The synopsis aims to understand the important concepts related to the research topic.

1. **Introduction**

The beginning of 21st century had a rapid development in the field of Computer Science especially in hardware. There has been a huge increase in the size of storage capacity of computer. The huge amount of data spread and created over the Internet, organizations, and institutes in the field of Medical Science, Business, Agriculture, Education, and E-commerce and so on, all over the world is in the process of continuous increment and has already surpassed terabytes. The explosive rate of data increment is growing everyday and the experts in the field have an opinion that the amount of information in world doubles every one or one and half year. Thus finding useful information from large amount of data is more difficult.

Statistical techniques play an important role to handle these types of datasets to find efficient solutions scientifically, so we can take decision for further movement. Data mining techniques are used to operate on large amounts of data to discover hidden patterns and relationships helpful in data mining. These rapid developments in computing and networking, along with e-commerce, can only increase interest in the theory and application of predictive analytics. Predictive analytics is an area of statistics that deals with extracting information from data and using it to predict trends and behavior patterns.

Often the unknown event of interest is in future, but predictive analytics can be applied to any type of unknown wheather it in the past, present or future [20] .

2. Review of Literature and Development in the subject

Data mining can be defined as an activity that extracts some new non-trivial information contained in large databases. The goal is to discover hidden patterns, unexpected trends or other subtle relationship in the data using combination of techniques from machine learning, statistics and database technologies [11]. These techniques facilitate the process of knowledge creation and contribute towards the further spread of data mining.

Data mining is an essential step in the knowledge discovery in databases (KDD) process which is the use of algorithms to extract the information and patterns by the KDD process. Data mining and knowledge discovery are interchangeably [9] more precisely, both terms are different. The term KDD refers to the automated process of knowledge discovery from databases, the process of KDD is comprised of many steps. Data mining is a step in the whole process of knowledge discovery which can be explained as process of extracting or mining knowledge from large amounts of data [2].

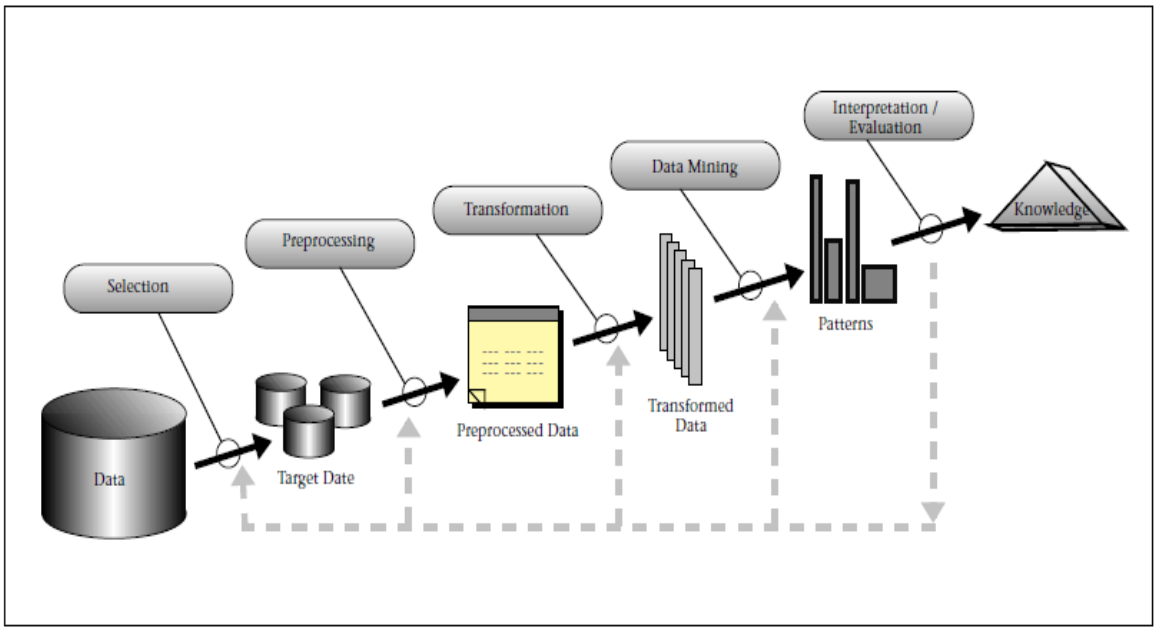

Figure 1 Data Mining and the KDD Process (Source: Fayyad, et .al., 1996)

Fig. 1 shows KDD process consists of iterative sequence methods as follows: Selection, Preprocessing, Transformation, Data mining, and Interpretation / Evaluation.

➢ **Data Mining Tasks:** The eight main functions of data mining are: Classification, Regression, Time series analysis, Prediction, Clustering, Summarization, Association Rule, sequence discovery which find sequential patterns in data[5].

➢ **Data Mining Techniques:** The most commonly used data mining techniques include artificial neural network, decision trees, classification and the nearest-neighbor method, each of which techniques analyzes data in its own and different ways: Cluster Analysis, Linear regression, Artificial Neural Networks (ANN), Genetic Algorithms, Rule induction (Association Rule), Nearest-neighbor method, Memory based reasoning, Logistic regression, Correlation, Discriminant analysis & Decision Tree, Factor analysis, Link Analysis, Visualization etc. It is possible that the research work can be carried out on the above mentioned & various different types of tools & techniques [5].

➢ **Statistical Data Mining**

Some of the Statistical Data Mining Techniques are as follows: Regression, Correlation, Generalized Linear Models, Analysis of Variance, Mixed-effect Models, Factor Analysis, Time Series Analysis, Cluster analysis [19] .

3. Objectives of the Study:

Some of the proposed objectives of the research study are stated below:

- To perform comprehensive analytical study of predictive techniques used in data mining.
- To perform comparative study of most common predictive techniques used in data mining.
- Try to develop new predictive model using various data mining techniques.
- To develop algorithm and mathematical/statistical model for such algorithms and to valide the developed model for various kind of data sets.
- To find new application of various predictive techniques in Data Mining.
- Try to develop new feature selection model.
- Try to find out the best predictive technique in Data Mining according to their applications.
- Try to develop new interdisciplinary approach of Data mining with Operations Research application areas to optimization the real life problems.

During the period of research many new objectives may be developed which will be beneficial for the society.

4. Methodology:

The research can be carried out by analysis of existing Data mining predictive statistical techniques on different datasets. The research requires various different kinds of databases, tools and techniques, data mining algorithms, scientific languages, different software and packages.

- **Databases**: Data Mining can be carried out on different types of databases: Relational Databases, Transaction Databases, Object -Oriented Databases, Deductive Databases, Spatial Databases, Temporal Databases, Multimedia Databases, Heterogeneous Databases, Active Databases, Legacy Databases and Internet Information Base Databases.
- **Data Mining Tools**: Most data mining tools can be classified into one of three categories: traditional data mining tools, dashboards, and text-mining tools.

- ***Traditional Data Mining Tools.*** Traditional data mining programs help companies establish data patterns and trends by using a number of complex algorithms and techniques.
- ***Dashboards.*** Installed in computers to monitor information in a database, dashboards reflect data changes and updates onscreen-often in the form of a chart or table enabling the user to see how the business is performing.
- ***Text-mining Tools.*** The third type of data mining tool sometimes is called a text-mining tool because of its ability to mine data from different kinds of text from Microsoft Word and Acrobat PDF documents to simple text files

- **Data Mining Algorithms:** The Data Mining algorithms are categorized under different types such as Clustering, Classification, Statistical Learning, Association analysis and Link Mining.
- **Scientific Languages:** The scientific languages used for developing the Data Mining applications are as categorized in two parts: (i) Open Source Languages (Python, R, Hadoop Based like Hive, Pig), (ii) Other Languages (Java, SAS)

- **Software and Packages:** The most used packages for Data Mining are: SPSS/ SPSS Clementine, Salford Systems CART/MARS/TreeNet/RF, Yale (now Rapid Miner), SAS / SAS Enterprise Miner, Angoss Knowledge Studio / Knowledge Seeker, KXEN, Weka, R, Microsoft SQL Server, MATLAB.

  The research methodology uses above mentioned different software packages to analyse and compare the predictive statistical techniques on different datasets.

5. Importance of Study/Society application:

- **International Status**

Hong and Weiss (2001) examined a few successful applications areas and their technical challenges. They review the key theoretical developments in PAC and statistical learning theory that have lead to the development of support vector machines and to the use of multiple models for increased predictive accuracy. In this way they presented on overview of some notable advances in predictive modeling. They were examining three challenging application areas: Insurance, Froud detection and Text categorization.

Olaiya and Adeyemo (2012) investigated the use of data mining techniques in forecasting maximum temperature, rainfall, evaporation &wind speed. This was carried out using ANN and DT algorithms and meteorological data was collected between 2000 and 2009 from the city of Ibadan, Nigeria. In this work the C5 decision tree classification algorithm was used to generate decision tres and rules for classifying weather parameters. The result show that, Data Mining techniques used for weather forecasting and climate change studies.

Joseph, Sadath and Rajan (2013) make comparative study on various techniques and methods. In their study, they focused on some of the most common data mining techniques and methods used in today day-to-day life and business predictions. They observed that, by using cluster analysis one can find out customers behavior viz. according to product or scheme activities in business and react to the marketing offers particular to their business. Also, Nearest-Neighbor technique is found to be easy to use and understand to predict value of the object based on its nearest neighbors.

Olafsson and Wu (2008) performed a survey of operations research and data mining and focused on both of the above mentioned intersections. They mentioned the use of Operations Reseach (O.R.) techniques may be useful in formulating and solving numerous data mining problems. The goal of their study is to illustrate the range of intersections of the two fields of OR and data mining.

- **National Status**

Beniwal and Arora (2012) provide a survey of various feature selection techniques and classification techniques used for mining. The data mining tasks can be broadly classifiesd in two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks

perform inference on the current data in order to make predictions. According to different goals, the mining task can be mainly divided into four types: class/concept description, association analysis, classification or prediction and clustering analysis.

Bhardwaj and Pal (2011) have used baysian classification method for student database to predict the students division on the basis of previous year database. To improve the division of the student this study is helpful for students and teachers. Also, work to identify those students which needed special attention to reduce failing ratio & taking appropriate action at right time. This study also shows that acadamic performance of the students is not always depending on their own efforts, other factors have got significant influence over student's performance. Vathsala H & S.G. Koolgudi (2016) presents a hybrid model to better predict Indian summer monsoon rainfall. In algorithm, they use three components: (i) feature selection (ii) dimensionality reduction and (iii) classification of features for predicting ISMR. Techniques used include closed itemset generation based association rule mining for feature selection, cluster membership function for dimensionality reduction and simple logistic function for classification. The study provides vital information to predicting rainfall into flood excess, normal deficit and drought.

Mohan and Ramsundram (2016) developed a model to analyze the historical reservoir operation database & recover the temporal relationship that exists between the reservoir release and other reservoir variables. Predictive data mining algorithms such as Artificial Neural Network (ANN) and Regression have been used for single reservoir operation data mining (SROD) modeling framework to explore the temporal dependence between different variables of reservoir operation. Thus predictive temporel data mining approach is very useful for evolving knowledge based reservoir operation rules.

- **Significance of the study**

Data Mining is widely used in diverse areas. Data mining has become not only important research area, but also one with large potential in the world.There are number of commercial data mining system available today yet there are many challenges in this field. The list of areas where data mining is widely used is as follows: Financial data analysis, Retail Industry, Telecommunication industry, Biological data analysis, Intrusion detection and other scientific applications.

Educational data mining is also used in many areas like computer science, education, psychology, psychometrics, statistics, intelligent tutoring systems, e-learning, adaptive

hypermedias etc. when analyzing large datasets in order to resolve educational research issue[1].

Temporal data mining has emerged as research area in rescent years for researchers all over the world in much different area. The field of temporal data mining is concerned with case of ordered data streams with temporal interdependencies [11].

Weather forecasting has been one of the most scientifically and technologically challenging problem around the world in the last century.To make an accurate prediction is one of the major challenges facing scientists all over the world. The list of areas where data mining techniques used is weather prediction and climate change, maximum temperature, minimum temperature, rainfall, evaporation and wind speed etc[15] .to solve these problems ANN and Decision Trees (DT) is helpful.

6. Proposed work Plan/ Formulation and Structure of Study:

- **Year wise Plan of work and targets to be achieved**

❖ **First Year**

1. To collect and study of review of literature of Data Mining.
2. Collection and Study of different Data Mining data sets.
3. Study of different Data Mining Software packages.

❖ **Second Year**

1. Try to develop statistical model for data mining.
2. Selection of one of the standard datasets for application of predictive data mining techniques for comparison of results obtained.
3. Detailed study and collaborative analysis of the results obtained.
4. To perform comparative study of the use of different Data Mining Software packages for the same data sets.

❖ **Third Year**

1. The publication of the findings of the research.
2. Publication of papers in various national and international journals & conferences.
3. Thesis writing.

7. References:

[1] Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. JEDM-Journal of Educational Data Mining, 1(1), 3-17.

[2] Beniwal, S., & Arora, J. (2012). Classification and feature selection techniques in data mining. International journal of engineering research & technology (ijert), 1(6).

[3] Bhardwaj B. K., Pal S. (2011). "Data Mining :A Prediction for performance improvement using classification", International Journal of Computer Science and Information Security, 9(4), 136-140.

[4] Chou, J. S., & Lee, C. M. (2014). Integrating the geographic information system and predictive data mining techniques to model effects of compound disasters in Taipei. Natural hazards, 70(2), 1385-1415.

[5] Dunham M. H.(2006). "Data Mining-Introductory and Advanced topics", Pearson.

[6] Güiza, F., Van Eyck, J., & Meyfroidt, G. (2013). Predictive data mining on monitoring data from the intensive care unit. Journal of clinical monitoring and computing, 27(4), 449-453.

[7] Hewett, R. (2003). Data mining for generating predictive models of local hydrology. Applied Intelligence, 19(3), 157-170.

[8] Hong, S. J., & Weiss, S. M. (2001). Advances in predictive models for data mining. Pattern Recognition Letters, 22(1), 55-61.

[9] Joseph, M. V., Sadath, L., & Rajan, V. (2013). Data Mining: A Comparative Study on Various Techniques and Methods. International Journal of Advanced Research in Computer Science and Software Engineering, 3(2), 106-113.

[10] Klionsky, D. M., Oreshko, N. I., & Geppener, V. V. (2011). Application of a Data Mining approach to the processing of telemetric signals. Pattern Recognition and Image Analysis, 21(4), 720-730.

[11] Laxman, S., & Sastry, P. S. (2006). A survey of temporal data mining. Sadhana, 31(2), 173-198.

[12] Mohan, S., & Ramsundram, N. (2016). Predictive Temporal Data-Mining Approach for Evolving Knowledge Based Reservoir Operation Rules. Water Resources Management, 30(10), 3315-3330.

[13] Oatley, G., Ewart, B., & Zeleznikow, J. (2006). Decision support systems for police: Lessons from the application of data mining techniques to "soft" forensic evidence. Artificial Intelligence and Law, 14(1-2), 35-100.

[14] Olafsson, S., Li, X., & Wu, S. (2008). Operations research and data mining. European Journal of Operational Research, 187(3), 1429-1448.

[15] Olaiya, F., & Adeyemo, A. B. (2012). Application of data mining techniques in weather prediction and climate change studies. International Journal of Information Engineering and Electronic Business, 4(1), 51.

[16] Singh, Y., & Chauhan, A. S. (2009). Neural networks in data mining. Journal of Theoretical and Applied Information Technology, 5(6), 36-42.

[17] Vathsala, H., & Koolagudi, S. G. (2016). Long-range prediction of Indian summer monsoon rainfall using data mining and statistical approaches. heoretical and Applied Climatology, 1-15.

[18] Veeramachaneni, K., Vladislavleva, E., & O'Reilly, U. M. (2012). Knowledge mining sensory evaluation data: genetic programming, statistical techniques, and swarm optimization. Genetic Programming and Evolvable Machines,13(1),103-133.

[19] http://www.tutorialdpoint.com/data_mining/dm_applications_trends_html

[20] http://www.wikipidia.com/data_mining/predictive analytics

Signature **Signature**

(Research Scholar) **(Research Guide)**
(Mr. Tangawade Atish Sitaram) **(Dr. Aniket A. Muley)**

Place: - Nanded
Date: -